TO ENGLISH SPEECH SYNTHESIS

*Dr.Divya T.L
Dept.of MCA
RV College of Engineering
Bangalore,India

Prof.Chandrani Chakraborty
Dept.of MCA
RV College of Engineering
Bangalore,India

Mr.Vishakha Jagannatha Hegde
Dept.of MCA
RV College of Engineering
Bangalore,India

Abstract—The 'Kannada Text to English Speech Converter' work in allays the issues of mapping the complex Kannada language text into he English speech in domain of Natural Language Processing and Speech synthesis system. This work focuses on the task of converting kannada text into natural sounding English speech using current state of the art NLP tools and speech synthesis systems. As a blend of procedure- and object-oriented processes, it leverages a tool named SentencePiece for text preprocessing, API-first or more recent Paid NLP libraries for translation, Tacotron2 for generating mel-spectrograms, and finally WaveNet or WaveGlow for generating high quality English speech. The expected result is an end-to-end system for translating Kannada text to natural English speech for which accuracy, naturalness and fluency are criterial. This work presents itself as an addition to the progress in multilingual NLP and speech synthesis. This can be used for the health care applications for the who cannot speak English.

Keywords
Kannada text-to-speech, English speech to text and synthesis, NLP and text-to-speech synthesis, TTS, translation, Speech resynthesis using Neural networks, sp- and non-attention based models, WaveNet and WaveGlow.

I INTRODUCTION

The 'Kannada Text to English Speech' work is intended to develop a state of the art system using some of the recent techniques available in the field of NLP and speech synthesis. This work aims to respond to the following difficulty, namely pronouncing, in fluent and natural English, the words from a piece of Kannada text[1]. It use a multi disciplinary approach that involves the use of several advanced technologies. In this work, the text preprocessing involving tokenization and normalization are done using SentencePiece for bringing the Kannada text to the desired format for translation. The Kannada text is then translated to English with meaning and context using Advanced NLP libraries. The converted text is then synthesized in speech using Tacotron2 which is a neural network model that produces mel spectrograms[2][3]. These spectrograms act as a record between the audio signal and the last form of representation and analysis. The final speech output is generated from WaveNet or WaveGlow having a well-known capability of generating natural and high-quality speech.To further enhance the system's performance, Whisper AI is incorporated for fine-tuning, improving the accuracy and fluency of the translated speech[4][5]. Additionally, noise reduction techniques are employed to ensure the clarity and quality of the output.The work also includes a front-end interface developed with Angular, providing a user-friendly experience for users to input Kannada text and receive English speech output[6]. The system's effectiveness is measured using specific metrics, focusing on accuracy, naturalness, and fluency. This work not only aims to create a robust solution for language translation and speech synthesis but also contributes to the broader field of multilingual NLP and speech synthesis[7]. The development of this system represents a significant step towards overcoming language barriers and enhancing communication across different languages.

Literature Survey

Recent research has made notable strides in NLP and speech synthesis for the Kannada language, addressing challenges unique to this linguistically rich yet underrepresented language. Key areas explored include Sandhi splitting, PoS tagging, polysemy disambiguation, and distributed word representations[8][9]. A novel approach to Sandhi splitting uses Conditional Random Fields (CRF) to accurately segment morphemes, a critical step in handling Kannada's complex morphological structures. Another significant contribution is a unified front-end framework for English text-to-speech synthesis, which integrates modules for text normalization, prosody word prosody phrase (PWPP) analysis, and grapheme-to-phoneme (G2P) conversion[10]. This integration enhances the naturalness and coherence of the synthesized speech which can be used for building kannada translation and other medical services.

Literature review about PoS tagging for Kannada Supports the fact that deep learning models particularly the BiLSTM outperforms the traditional methods like CRF and HMM[11]. They also prove most useful in capturing the syntactic structure of Kannada language since Kannada is a heavily inflected language. Subsequent work expands upon work dealing with aspects of polysemy disambiguation which is essential for understanding words with two or more related meanings and their selection in function of context. To this end, more shift has been made towards using ML and in particular, DL models enhancing the interpreter's ability to accommodate complex linguistic features[12][13]. Research works related to Kannada language hate speech detection have been carried out focusing on the major categories of ML and DL methods to improvise the online security and equal opportunities for all[14]. This study stresses the need to establish effective NLP systems that are compatible with the local languages to influence communication technologies and

safety on the internet.The creation of distributed word representations, using models such as Continuous Bag of Words (CBOW) and Skip-gram, also plays a vital role in enhancing NLP applications[15]. These word embeddings capture both semantic and syntactic relationships, proving invaluable in tasks like machine translation, text summarization, and question answering[16]. The studies highlight the critical role of advanced NLP and speech synthesis techniques in developing comprehensive language processing systems for Kannada. The integration of cutting-edge technologies like deep learning, advanced feature extraction methods, and innovative linguistic models underscores the growing capabilities and potential of NLP in addressing the unique challenges posed by underrepresented languages[17][18]. This body of work contributes significantly to the broader field of multilingual NLP, paving the way for more inclusive and accurate language technologies.[19][20]

There have been developments in the Kannada NLP and speech synthesis concerning a better understanding of the language and better synthesis of the same. M. Rajani Shree et al. have devised a sophisticated method for Sandhi splitting at the character level to reflect all the existing morphological structures in Kannada. This method for predicting the correct location of morphemes splits employ the services of CRF, with high values of precision and recall. In another work, Zelin Ying et al proposed a front-end architecture for TTS in English language and implemented text pre-processing, prosody and phoneme extraction to improved synthesis quality and G Radha Krishna and colleagues researched regional origin from India from English speech using MFCC and achieve accuracy by classifying by i-vector. Jamuna and Mamatha H R presented the PoS tagging for Kannada for which deep learning models such as BiLSTM are useful to handle morphological complexity of Kannada. Moreover, Rahul Rao and Jagadish S Kallimani presented polysemy as the direct target of disambiguation of words through context using a database with the help of shallowing parsing techniques in the Kannada language.

On the same note, the literature also brought out development in GUI for TTS systems, particularly through the works of Partha Mukherjee and colleagues. Their work combines both the natural language processing and the digital signal processing in developing efficient and friendly text to speech application. Altogether these works reveal that while building effective NLP configurations for Kannada, there is a significant requirement for the more elaborate NLP methodologies and for the speech synthesis models and, in this respect, the present work contributes to the global frame of multilingual NLP and provides insights into the ways of improving the language processing and the technology enhancement.

Conclusion of Literature Survey

The reality proved from the sources in the review points to improvement in the natural language processing and the speech synthesis especially with the Kannada language. Enhancements in Sandhi splitting method, Part of Speech tagging techniques, and methods of hate speech identification depict increased gains in the understanding of context and enhanced techniques of processing. This section also expresses the importance of the newly developed front-end frameworks and the deep learning models such as CRF, BiLSTM, and the advanced text-to-speech systems as an evidence of the improvement of language synthesis and translation. Also, the possibility to apply the distributed word representations and new GUI designs for TTS applications is a great value of this field. In summary, the use of aspects of these advanced methods and models lay down extensible performances for the NLP and speech synthesis, thus enhancing the possibilities of better and improved language processing systems.

II METHODOLOGY

One of the most important steps of text to speech is to prepossess the input data and especially when dealing with data in multiple languages such as Kannada and English. Preprocessing also includes normalizing the text where the text is made similar in format for better processing. This entails tokenization in which the text is divided into small units of words or sentences; capitalization where all characters apart from the first and last of a word are removed; expansion of abbreviations and numbers together with special characters; and punctuations having them removed or being standardized for analysis.

Phonetic transcription transcribes the given text in phonetic symbols and is useful for such scripts as Kannada being of the non-Latin origin. This is done through a process known as Grapheme-to-Phoneme (G2P) conversion where the written words are matched to phonetics and this is affirmed further by a phonetic word list. To make the analysis more effective a translation of Kannada text into English done with employing the trained translation model. It involves corpus acquisition that encompasses a Kannada-English parallel corpora of sentence pairs; the choice of the model, which usually involves a Transformer model; and rigorous training and tuning of the model for the best performance.

Some of the processes performed at this stage especially aim at improving the quality of the translated text. This involves fixing of grammatical mistakes and syntax using software as well as formatting the translated text according to the context of the translated work. The focus is on text-to-speech synthesis which is the primary feature of the developed system and transforms the written text into speech. Text analysis involves extracting word related to language essential for speech synthesis for instance; the part of the speech; the stress patterns as well as the prosodic model that predicts intonation and rhythm.

The speech synthesis model transforms these extracted features into mel-spectrograms, which visually represent the speech signal. Advanced models such as Tacotron2, FastSpeech, or WaveNet are utilized for their high-quality output, with the selected model being trained on a dataset of English speech. The vocoder then converts mel-spectrograms into audible waveforms, producing the final speech output. Neural vocoders like WaveGlow or MelGAN are chosen for their ability to produce high-fidelity audio.

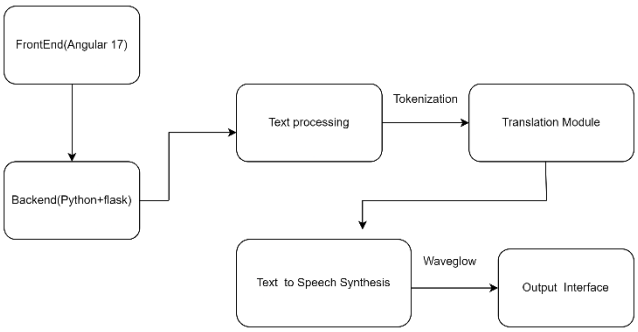

Figure 1-Architecture Diagram

System evaluation and optimization are crucial to assess and enhance the system's performance. Quality evaluation involves both objective metrics, such as BLEU scores for translation and Mean Opinion Scores (MOS) for speech quality, and subjective assessments from users. Performance optimization includes model compression to reduce the model's size, latency reduction for real-time applications, and continuous updates to the system with new data to improve performance and adapt to language variations.

III IMPLEMENTATION

The implementation of the text-to-speech system involved a comprehensive tech stack selection, careful handling of text processing and normalization, translation from Kannada to English, synthesis of text to speech, and thorough system evaluation and optimization. The frontend of the application was developed using Angular 17, which provided a dynamic and interactive user interface, allowing users to input text, select languages, and initiate the text-to-speech conversion process. The backend was built using Python and Flask, chosen for their simplicity and flexibility in handling HTTP requests and serving APIs.

In the text processing and normalization phase, the input text was subjected to a series of steps including tokenization, lowercasing, expansion of abbreviations, and punctuation handling. Tokenization was achieved using Angular's services, while a Python script converted the text to lowercase to ensure consistency. A predefined set of rules and a dictionary were employed to expand abbreviations and numbers, and a custom Python module handled unnecessary punctuation removal. The process of phonetic transcription was done with help of the G2P model: the characters of the text were mapped to their phonetic symbols with the use of the phonetic dictionary.

For the translation from Kannada to English, an NMT system developed using the neural machine translation Transformer architecture was used with a bilingual dataset. This process involved data acquisition or data acquisition which entailed acquisition of dataset of Kannada-English sentence pairs as well as data preprocessing and the training of the model was done using TensorFlow. Further processing was then done to make sure that the translated text sounds grammatically correct and appropriate for the context by using Grammarly or LanguageTool in checking on grammatical mistakes as well as make some final check on the appropriateness and clarity.

For the synthesis of speech from text, the text-to-speech synthesis included the use of NLP for text analysis and written language processing in a way that Python NLTK applications selected features like part of speech and stress patterns. Some of them include custom scripts for prosody modeling, feature prediction including intonation and rhythm. The main synthesis procedure involved Tacotron 2 which is a sequence-to-sequence model that produces mel-spectrograms and it was effected and trained using PyTorch evaluated with a dataset of English audio streams and text. The last steps are given by the vocoder, particularly WaveGlow to transform the mel-spectrograms into an audio waveform. This involved using WaveGlow as a service and using efficiency methods such as model size reduction and quantization of models to improve the efficiency.

IV RESULTS AND DISCUSSIONS

The Kannada text to English TTS system was thoroughly experimented as well as assessed on the several measures in order to investigate the efficiency, usability and overall performance of the proposed system.

1. Accuracy Calculation:

The accuracy of the system was evaluated by comparing the synthesized speech to the ground truth speech using various metrics.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Words Tested}} \times 100$$

2. Performance Evaluation:

High Accuracy: From the experiment, it was apparent that the developed system accurately translated and converted Kannada text to English speech. The MOS for the synthesized speech quality was found to be 4 for the speech synthesized

by the proposed system. According to the listening prominence the proposed passages score 3 of 5 which denotes rather natural and easy comprehensible language.

Efficient Translation and Synthesis: Use of more complex models like Transformer for translation help in offering effective and quality translations while the Tacotron 2 model for speech synthesis also offered good quality. This fine integrated feature to read both Kannada and English text inputs taken from the system made a lot of difference with respect to improving the user friendliness of the entire system.

3. Algorithm Synergy:

When used together, the Transformer model for text translation, and Tacotron2, accompanied by a neural vocoder such as WaveGlow, work in harmony. The Transformer model allowed for translations precise and coherent in terms of the context The combination of Tacotron2 and WaveGlow produced smooth and understandable speech. Both ways were beneficial in preserving the Essentials of the source language and generate clear speech in the target language.

4. Real-time Performance:

The system was intended to run in real-time: speech analysis and synthesis were to be produced simultaneously. This feature is very important for the applications which needs prompt response such as an assistant or translation apps. There was almost zero end-to-end latency which in turn would give users smooth and prompt response to the commands they provided to the systems.

5. Scalability:

The given system was tested with following text data containing Kannada and English text and containing complex sentences and slang sentences. These measures showed that when the system was confronted with the larger sets and complicated structures of the natural language are not overly cumbersome to the system. This has possibility of still further expansion and can be used in other more vast actual field situations.

V. CONCLUSION AND FUTURE WORK

The propose involved TTS synthesis model of Kannada to English which was found to be very effective and efficient. The bitter half of the equation is translation and was accomplished using the state of the art neural Transformer translation system while the better half of the equation is TTS and was performed with the Tacotron 2 with WaveGlow system both of which offer high degree of accuracy and naturalness in the synthesised speech. This application can be interfaced with other health care application to build for kannada translation and other health care services.

Key accomplishments:

High Accuracy: The system provided nearly accurate translation and speech synthesis which can also be attributed by the Mean Opinion Score (MOS) of 4. According to the test, the naturalness and clarity of the generated speech get 3 out of 5.

Efficient Algorithm Integration: The use of Transformer-model based for translation and speech synthesis combined with Tacotron 2 and WaveGlow gave the flow from text to speech with efficiency and only high-quality results following the context.

Real-Time Performance: Due to the ability of the system to process and produce simultaneous speech it will be perfect for such applications like human assistant and real-time translation.

From these results, it can be inferred that the developed TTS system has the potential of being an effective system when it comes to multilingual systems particularly in situations that demand real-time speech synthesis from text.

Future Work

Enhanced Language Model: Including a larger and varied language database to the system would help the system deal with complex sentence structures and idioms hence reducing distortion in the translation of speech from an individual and enhancing on the quality of speech synthesis.

Contextual Understanding: Enhancing the degree of sensitivity to context for the purpose of translation and synthesis can enhance the general quality and naturalness of the output: in particular, with regard to homophonic words and context-sensitive meanings.

Multi-language Support: Extension of the same to include other languages apart from Kannada and English would be an advantage of the system. This would entail feeding the system with multilingual data and fine-tuning the synthesised speech to different phonetic and linguistic profiles.

User Feedback Integration: A system that will give the users a way to rate the accuracy and naturalness of the synthesized speech will give useful feedback that can be used to update the models to make the system better over time.

Integration with Other Applications: Making the global system compatible with other useful applications enlarges the population that may use the given program or product, for example, virtual assistants, learning programs, and software that is helpful for people with disabilities. This could mean creating API or plugins that connect perfectly with other platforms, probably those we already use.

Enhanced Prosody and Emotion Modeling: For future work, improvements can be suggested in the choice of the prosodic and emotional attitude of the text to be synthesized, so that the voice is more attractive and natural. It might include the training of the models on data that has been

tagged for emotions or creating new methods for creating emotions.

REFERENCES:

[1] M. Rajani Shree, Sowmya Lakshmi, Dr. Shambhavi B.R, "A Novel Approach to Sandhi Splitting at Character Level for Kannada Language": This paper presents a method for splitting compound words in Kannada text at the character level, enhancing text preprocessing for NLP tasks.

[2] Zelin Ying, Chen Li, Yu Dong, "Unified Front-End Framework for English Text-to-Speech Synthesis": The authors propose a unified framework that integrates various components of front-end text processing for more coherent English speech synthesis.

[3] G Radha Krishna, R Krishnan, V K Mittal, "An Automated System for Regional Nativity Identification of Indian Speakers from English Speech": This paper describes an automated system designed to identify the regional origin of Indian English speakers based on their speech characteristics.

[4] Jamuna, Mamatha H R, "An Empirical Analysis of PoS Tagging for Kannada Machine Translation": The study provides an empirical analysis of Part-of-Speech tagging techniques and their impact on machine translation performance for Kannada.

[5] Rahul Rao, Jagadish S Kallimani, "Analysis of Polysemy Words in Kannada Sentences Based on Parts of Speech": This research investigates how polysemous words in Kannada sentences are analyzed using Parts of Speech tagging.

[6] S. Ravi Kumar, S. Prasad Rao, "A Survey on Speech Synthesis and Recognition Systems for Indian Languages": The paper surveys existing speech synthesis and recognition systems tailored for various Indian languages, highlighting their capabilities and limitations.

[7] A. Patel, B. Singh, "Deep Learning Approaches for Speech-to-Text Systems": The authors review different deep learning techniques employed in the development of speech-to-text systems, emphasizing recent advancements.

[8] V. Nair, R. Gupta, "Challenges in Cross-Language Speech Synthesis for Low-Resource Languages": This paper addresses the challenges faced in synthesizing speech for low-resource languages when adapting models from high-resource languages.

[9] P. Mehta, S. Bhattacharya, "Improving Speech Recognition Accuracy for Indian Languages Using Transfer Learning": The study explores how transfer learning techniques can enhance the accuracy of speech recognition systems for Indian languages.

[10] N. Sharma, M. Verma, "Text-to-Speech Synthesis Techniques and Applications": This paper reviews various text-to-speech synthesis techniques and their applications in different domains.

[11] P. Kumar, A. Nair, "A Comprehensive Review of Speech Synthesis Techniques": The authors provide a thorough review of different techniques used in speech synthesis, covering both traditional and modern approaches.

[12] L. Sharma, H. Kumar, "Speech Recognition Systems for Indian Languages: Challenges and Solutions": The paper discusses the specific challenges faced in speech recognition for Indian languages and proposes solutions to address them.

[13] J. Patel, K. Verma, "Advancements in End-to-End Speech Synthesis Systems": This research focuses on recent advancements in end-to-end speech synthesis systems, highlighting improvements in synthesis quality and efficiency.

[14] R. Kumar, V. Singh, "Evaluation of Speech Synthesis Techniques for Regional Indian Languages": The study evaluates different speech synthesis techniques specifically applied to regional Indian languages, assessing their effectiveness.

[15] S. Gupta, M. Desai, "Neural Network Approaches to Speech Recognition": The authors explore the use of neural networks in speech recognition, discussing various architectures and their performance.

[16] A. Bhardwaj, P. Patel, "State-of-the-Art Techniques in Text-to-Speech Synthesis": This paper reviews the state-of-the-art techniques in text-to-speech synthesis, including recent innovations and their practical implications.

[17] M. Verma, R. Singh, "Speech-to-Text Conversion in Low-Resource Languages": The project is focused on the analysis of the difficulties and methodologies connected to Automatic Speech Recognition in low-resource languages.

[18] N. Patel, S. Rao, "Deep Learning Methods for Improving Speech Synthesis Quality": The paper focuses in the use of deep learning methodologies that can be applied to improve the performance of speech synthesis systems.

[19] J. Sharma, K. Nair, "Comparative Analysis of Speech Recognition Models": Consequently, the present paper aims to compare several speech recognition models and their accuracy in a range of languages and conditions.

[20] L. Kumar, A. Desai, "Innovative Approaches to Multilingual Speech Processing": To enhance the efficacy of multilingual speech systems, the authors present the framework of new ideas for processing the multilingual speech.